A Part of Speech Tagset for Sinhala

This document summarises the issues identified in the discussions held while preparing the part of speech tagset for Sinhala. The tags identified are given at end of the document.

Issues

1. **Word Separation**

   In Sinhala, there is no generally accepted set of rules for word separation. But there seem to be groups of writers who conform to certain rules. This fact is reflected in the 10 million word Sinhala corpus too. A decision has to be taken to preserve the consistency of the part of speech tagging process.

   1.1. Separation of Particles and Post-positions

   E.g. ගියේය / ගියේ ය, ඔහුද මමද, ඔහුගේ, ඔහු ගේ, ඒ ගැනම

   1.2. Separation of Compound Nouns

   E.g. විශ්ව විද්‍යාලය

2. **Segmentation**

   It is the practice in Sinhala writing to combine two words belonging to certain grammatical categories to form one orthographic word.

   E.g. ලස්සනයි, විචාළෙත්, විසිනි, බැවිනි, හෙයිනි, හැකිද්දැයි, ඇතැයි, යැයි

3. **Multiword**

   Certain word combinations/phrases can function as one grammatical category.

   E.g. යටත් පිරිසෙයින්, කරණ කොට, යළි යළිත්, බල බලා, නට නට

4. **Unknown Grammatical Categories**

   It is hard to determine the grammatical categories of certain words as they do not show the characteristics of any known grammatical category by themselves, particularly in the initial part of some compound verbs.

   E.g. *මිය* ගියේය, *අසු* වී, *පත්* වූහ, *අවදි* වූයේ, *පහළ* වීමට

5. **Membership criteria for post-positions and particles of *Nipatha***

   By definition a post-position follows a noun or a noun phrase. In traditional grammar of Sinhala called them as *nipatha*. But in some cases *nipatha* follow too. In such context it is more appropriate to label them as particles.

   E.g. පොත බලන තෙක් / විට / කල්හි/ අයුරු / නිසා
   මගේ පොත ගැන / තුළ / නිසා

6. **Pronouns**

   One part of speech label can be assigned to all the pronouns. A detailed tag can be assigned simple table look-up by maintaining a list of all the pronouns as they form closed class.

TAG	Description	Example
NNR	Common Noun Root	මිනිස්, පුටු, බලු
NNM	Common Noun Masculine	මිනිසා, බල්ලා, ශිෂ්‍යයෝ, එළුවන්
NNF	Common Noun Feminine	නිළියෝ , ඇතින්න
NNN	Common Noun Neuter	පුටු, ගම
NNPA	Proper Noun Animate	රත්නායකගේ
NNPI	Proper Noun Inanimate	රත්නපුරේ, ලක්ස්ප්‍රේ
PRPM	Pronoun Masculine	ඔහු, ඒකා, කොයිකා
PRPF	Pronoun Feminine	ඇය, ඕකී
PRPN	Pronoun Neuter	එය, ඕක
PRPC	Pronoun Common	මම, ඔවුහු
QFNUM	Number Quantifier	එක, දෙවනි
DET	Determiner	මේ, ඒ, අර, ඔය, බොහෝ, සියලු,
JJ	Adjective	රළු, සුමුදු
RB	Adverb	වහා, ඉතින්
RP	Particle	ම, ලු, ය, ද, නම්, වැනි, වූකලී
VFM	Verb Finite Main	බලයි, බැලූහ
VNF	Verb Non Finite	බලා, බලමින්, බලද්දී, බලතොත්
VP1	Verb Participle 1	බලන,බැලූ, කළ
VP2	Verb Participle 2	බලනු, බැලුවා
VP3	Verb Participle 3	බැලිය
VP4	Verb Participle 4	බලන්නේ, බැලුවේ, කළේ
VNN	Verbal Non Finite Noun	බැලීම්, බැලිලි, බැලුම්
POST	Postpositions	ගැන, ලෙස, සඳහා
CC	Conjunctions	සහ , ද, සමඟ, හෝ
NVB	Noun in Kriya Mula	පාඩම් කරනවා, භාජනය කරනවා
JVB	Adjective in Kriya Mula	කීකරු වෙනවා, එකග වෙනවා, අඩු කරනවා
UH	Interjection	අහෝ , චිඃ, ෂික්, ආඃ
FRW	Foreign Word	Computer
SYM	Not Classified	A4